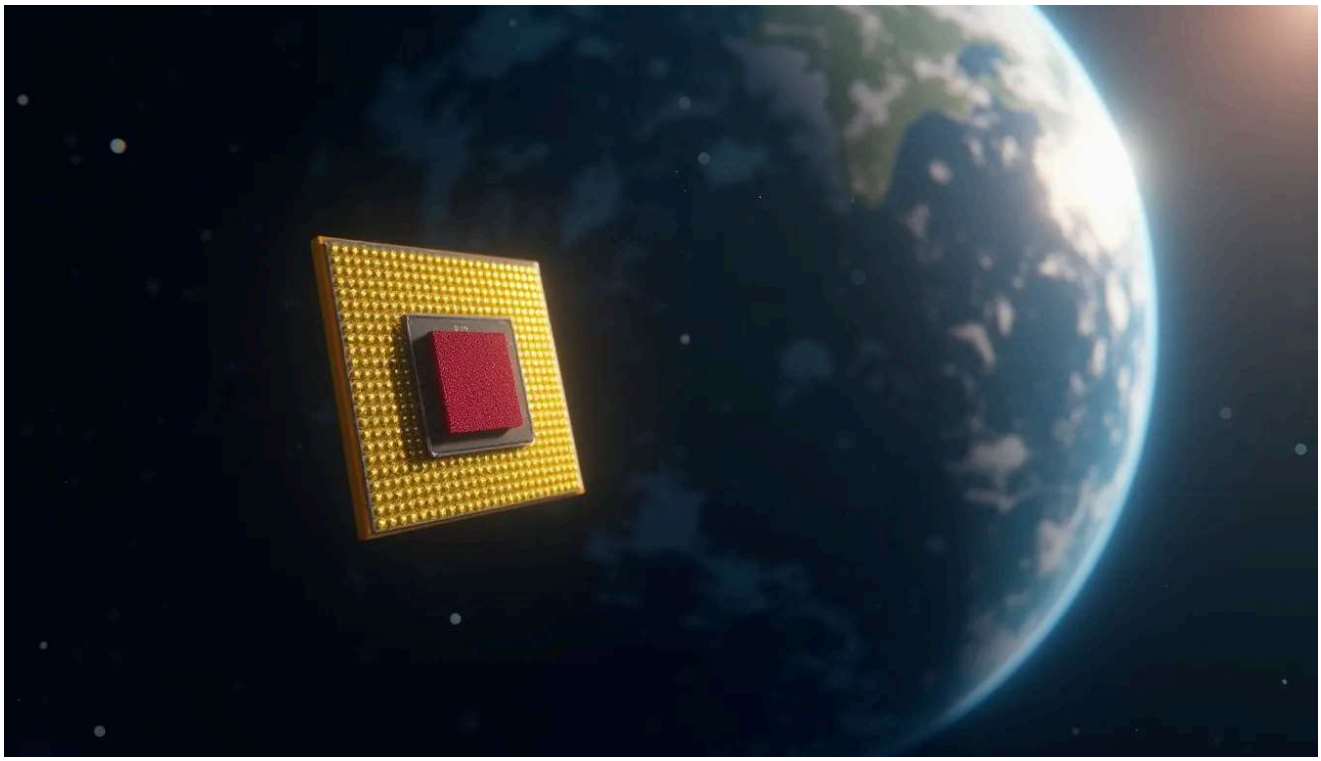
January 2025

DeepSeek might be hiding its hardware, Samsung and Micron refuse to give up, and the RTX 5090 stuns the industry.

HITESH KUMAR

JAN 31, 2025

Share



A person comparing their own size to the earth would have an easier time comprehending that size difference than trying to compare the size of a transistor to the chip it has been etched into.

This month's updates:

- DeepSeek's hardware inventory possibly much greater than reported
- Samsung and Micron still competing for HBM, fighting for Nvidia's GPUs
- Nvidia's GB200: Mixed sentiment from industry on persistent overheating reports
- Qualcomm and Arm to compete again: The datacentre Arm-based CPU market

- The incredible engineering of the world's best gaming GPU — the Nvidia RTX 5090

Vendor spotlight:

- GigaIO

One-pagers:

- Overclocking
- Intel AVX/AMX

This month's updates:

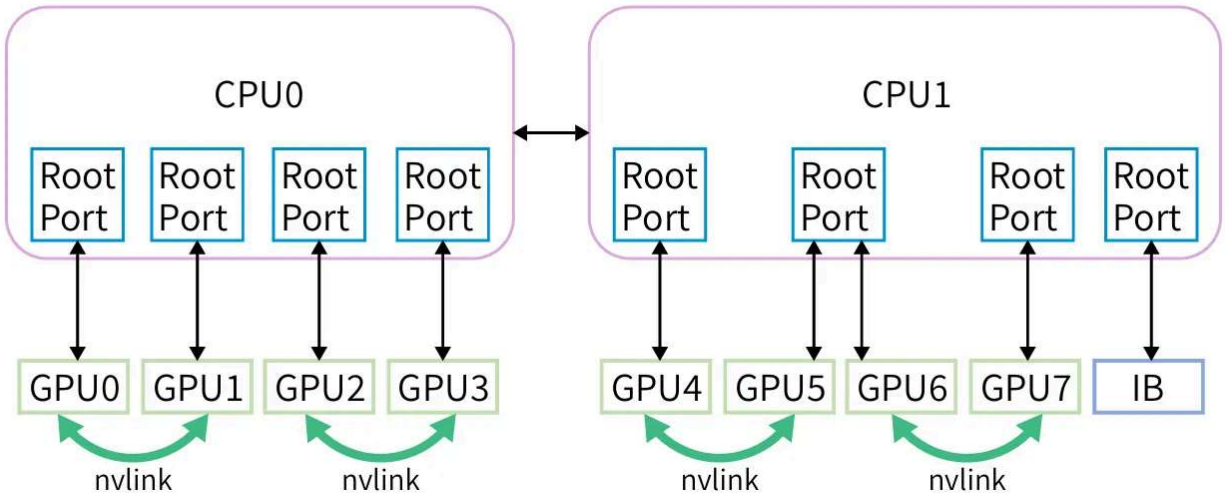
DeepSeek possibly under-reports AI hardware infrastructure

The AI model that wiped an estimated $1.5 Trillion in value off global stocks was not an automated trading or banking service gone wrong, but just a free to use - and very capable - chatbot. DeepSeek's R1 was reportedly trained on just 2048 H800 GPUs, but some sources claim that significantly more compute was used.

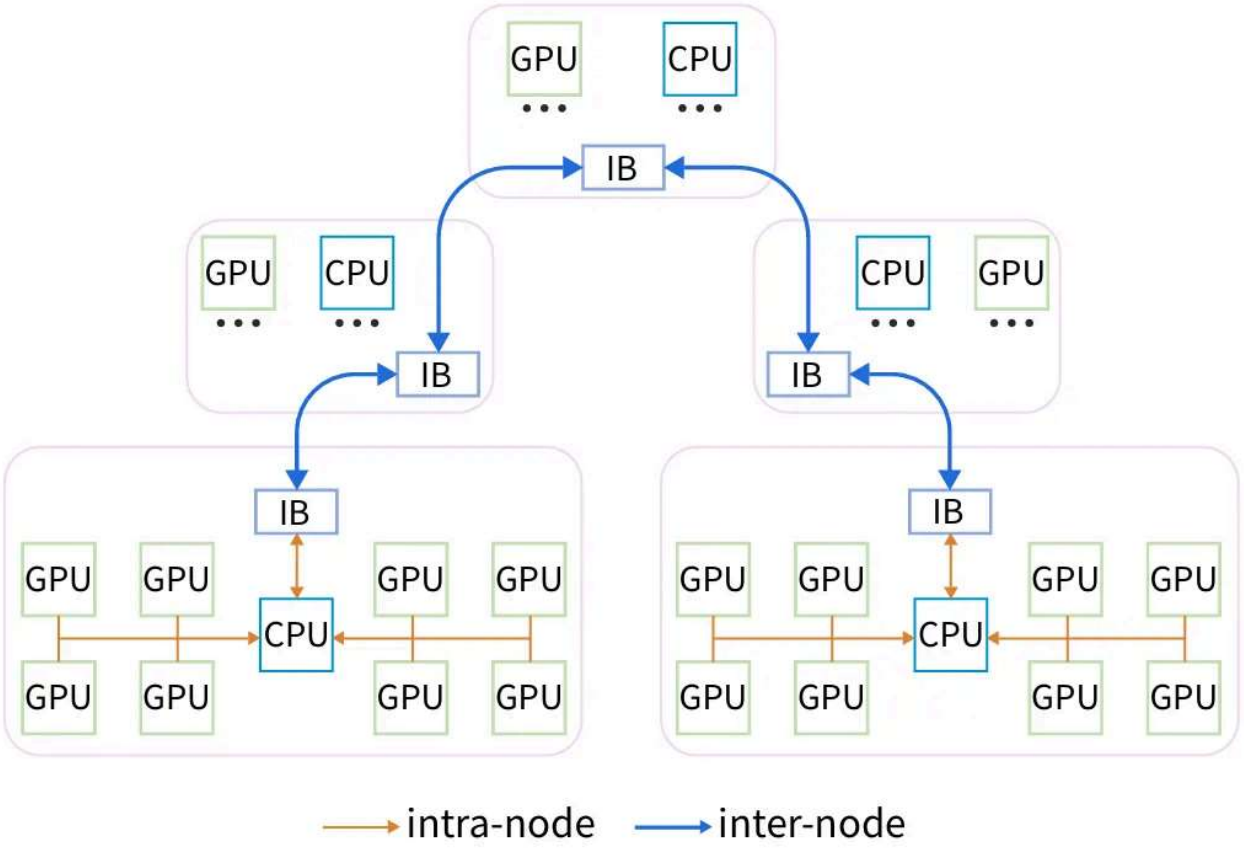
Though the paper claims that only 2048 Nvidia H800 GPUs and 5.6M USD were used during model training, speculation and rumours from industry have brought the legitimacy of this statement into doubt. DeepSeek (owned by Chinese hedge fund "High-flyer") open-sourced their free-to-use "R1" model and shocked the market by quickly reaching second place on various public leaderboards and benchmarks, just short of OpenAI's o1. What's more, the model is significantly smaller and cheaper to run than its American/European counterparts and is freely available for anyone to build upon and run locally. But many sources claim it's likely that DeepSeek used a lot more compute than they report.

Huida (Nvidia's trading name in the Chinese market) designed the H800 AI accelerator in response to the 2022 US chips act, with it being essentially a reduced bandwidth H100 tailored to avoid export restrictions and meet the high demand for AI training compute from restricted countries. Both accelerators are

almost equivalent except for the H800 supporting only 8 x 50GB/s NVLink channels instead of the 18 available on H100s. ScaleAI CEO Alexandr Wang states that industry rumours hint at DeepSeek owning over 50,000 H100 GPUs, likely acquired through Singapore to circumvent trade restrictions, and SemiAnalysis stating that its more likely about 10,000 each of the H800 and H100s, with 30,000 H20s



Based on their "Fire-flyer" paper, it's evident that DeepSeek staff have extensive experience in optimising and maintaining infrastructure containing thousands of Nvidia GPUs with InfiniBand fabrics. The paper outlines how they designed and benchmarked a cluster of ~10,000 A100 GPUs, significantly deviating from Nvidia reference architecture for large cost reductions with relatively small performance drops. Most notably, their use of Connecting GPUs and IB NICs to CPUs via PCIe ports directly with no switching and developing their own collective communications library "HFReduce" which outperforms NCLL on their P2P NVLink setup, are both unique and very resourceful initiatives.



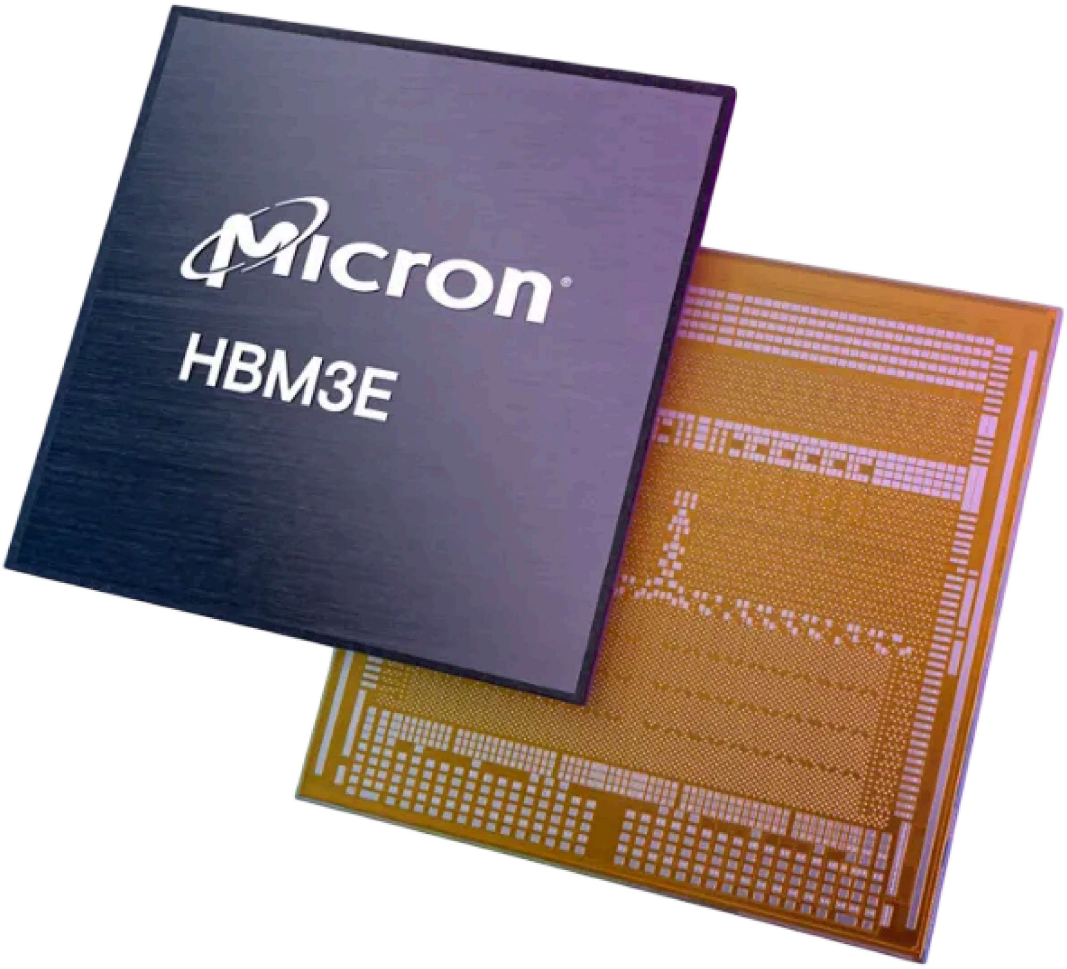
Source: Highflyer

Samsung and Micron still in the race for HBM

Last quarter, SK-Hynix and Nvidia both confirmed plans for 6th gen. HBM4 with the memory manufacturer agreeing to meet the 1H2025 deadline for sampling their devices on Nvidia's "Rubin" architectures. Now, Samsung and Micron both formally announce that they too are still competing for the future of this market.

In mid-4Q24, Nvidia requested SK-Hynix to accelerate their development of HBM4 devices, preparing samples for testing and validation on Nvidia's "Rubin" prototypes due in 2Q25. This resulted in them bringing forward their roadmap by 6 months, setting the date for mass production to the end of 3Q25, a timeline that originally had not been shared by their competitors.

Now, to regain their shrinking share of the HBM market, Samsung announced that they too will be releasing HBM4 samples for the end of 2Q25, likely to compete for the same trials on Nvidia hardware. It's unclear whether Samsung will meet this deadline due to recent heat/power management issues with their (4^{th} and 5^{th} gen.) implementations, with Nvidia repeatedly failing Samsung's attempts at passing validation. Their specific HBM4 designs also reportedly include using world-leading 1c 10nm DRAM dies, with logic dies being incorporated into the device for multiple, smarter buffers and providing some compute-in-memory capabilities. On top of this, Samsung also have declared intent to develop custom HBM4 devices for chip manufacturers such as Microsoft and Meta who are likely to be in the market for memory with narrower bus widths and higher overall density.



In addition, Micron, the smallest of the 3 major memory manufacturers has announced that they are conducting final tests for their implementation of high-density 16hi HBM3e, with mass production due before 2026. Despite being behind SK-Hynix in 16hi, Micron might well have a better chance than Samsung of increasing their HBM market share over the next year due to them famously winning the Nvidia H200 contract for 8hi HBM3e and hence being able to make significant investments in expanding their production capacity over the coming years. Unlike the other two, Micron keeps its HBM4 mass production plans to 2026/27.

Mixed market sentiment on Nvidia's GB200 rack-scale SKUs

Nvidia's B200 GPUs might possibly be at the centre of another overheating issue, this time stemming from the extreme density of their rack-scale SKUs, the NVL36/72s. Taiwanese suppliers such as Foxconn, Winstron, and QCT amongst

the largest, claim that these rumours are incorrect and that orders are being fulfilled as expected.

From as far back as mid-3Q24, rumours have been circulating of Nvidia's "Blackwell" rack SKUs experiencing overheating issues, causing issues for major customers such as Microsoft and Amazon. Recently, these reports have resurfaced as reductions in order volumes and an increase in demand for older "Hopper" H1/200 GPUs have been confirmed, with OpenAI reportedly asking Microsoft to provide more capacity on H100s to make up for the delays.

Source: Nvidia

With B200s having a TDP of up to 1000W and Grace CPUs designed for 500W, a single 1RU GB200 tray with 2 Grace and 4 Blackwell chips is estimated to have a maximum power draw of ~6300W under sustained high utilisation, requiring advanced direct liquid cooling solutions that very few datacentres in the world have the infrastructure for. In total, an NVL72 rack holds 18 such compute trays

as well as 9x1RU NVswitch trays for the GPU-GPU data fabric, all connected with active copper cabling. Whilst a variety of OEMs have now released their own implementations of the liquid cooled GB200 NVL72, it's likely that testing outside of curated lab environments has been rare.

Source: Nvidia

Taiwanese manufacturers and integrators such as Foxconn, Winstron, and QCT all deny these rumours, reporting that the current shipments are on schedule and are not affected by overheating issues. Instead, analysts cite cost vs benefit calculations compared to the tried and tested H1/200 HGX servers that are widely available to many buyers, as well as the timeline for Nvidia's B300 (previously named the B200 Ultra) and "Rubin" series GPUs being close enough to affect ordering times for hyperscalers.

Qualcomm and Arm to compete over datacentre Arm CPUs

Another established mobile chipmaker now moving into the datacentre CPU space, Qualcomm has been hiring for server design roles recently. At the same time Arm may be preparing for an acquisition of the high-performing Arm CPU designer, Ampere.

With Nvidia's Grace chips having shown success in the Arm server CPU market - albeit likely due to them being packaged in the GB200 rack-scale SKUs - and Apple showing signs of entering the same space with names such as Broadcom and Foxconn, it was always expected that other semiconductor design teams

would follow suit. Now, both Qualcomm and Arm appear to be taking significant actions towards this, with aggressive hiring and discussions of acquisitions.

Qualcomm, having previously forayed into this space in 2017 with their Centriq 2400 series, couldn't gain significant traction in the market due to the dominance of x86-based competitors and the difficulty of porting many applications to Arm architectures. Now, preferences and capabilities in industry have shifted towards energy efficiency, evidenced by a movement towards names such as Nvidia's grace, Ampere's Altra, and AWS's Graviton to name a few. To confirm their intent, Qualcomm have hired Intel's server CPU lead and have been openly looking for other CPU and server design roles.

At the same time, Arm have reportedly entered discussions along with SoftBank on acquiring Ampere, the market leader now in high performance, high core-count ARM server CPU. Until now, Arm has kept to developing and licensing semiconductor IP such as their individual cores for mobile and server applications (the Cortex and Neoverse series) which have seen success in processors made by Nvidia, AWS and Google to name a few. Ampere, on the other hand develop very high-core count processors such as their 96-192 core "AmpereOne" M and MX series and their upcoming "Aurora" 512-core SKU scheduled for 2025. Some of the concerns that regulators might have include possible conflicts between existing licensing agreements, Oracle's 29% share in Ampere, and Ampere's own plans for an IPO which are still underway.

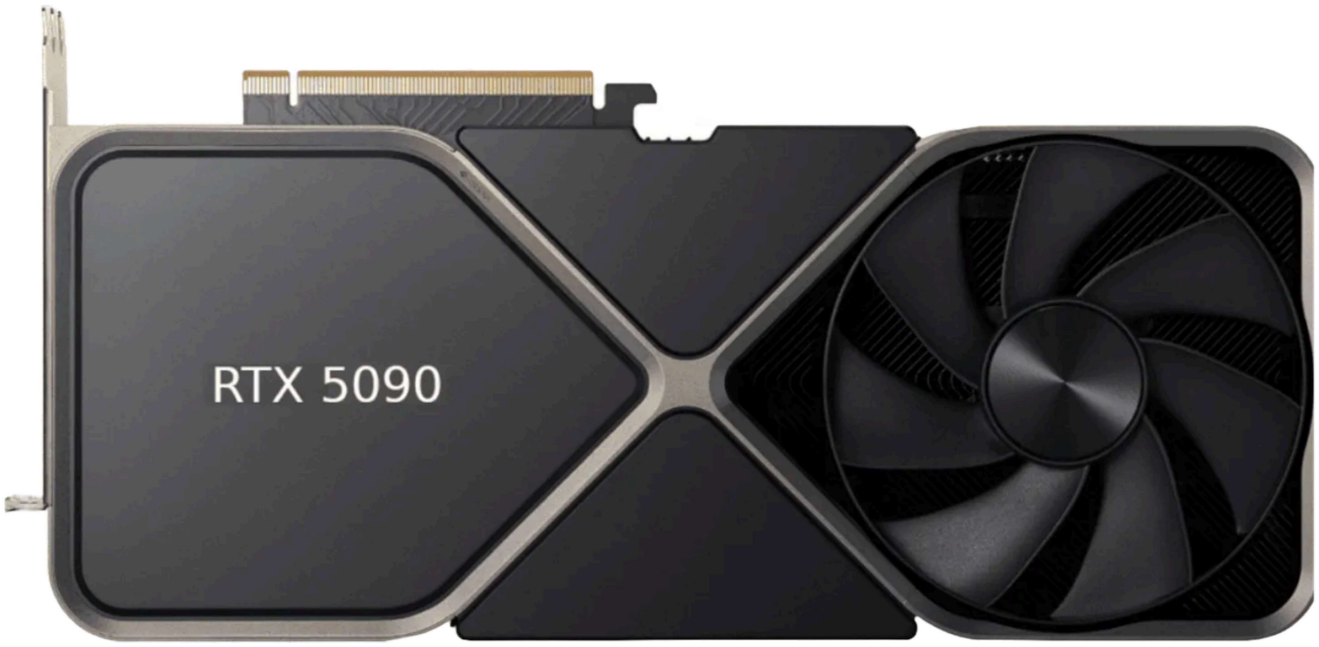
The incredible engineering of Nvidia's RTX 5090

Nvidia showcases their expertise in the gaming GPU space with their latest series of RTX graphics card built on their new "Blackwell" architecture. For the 5090 - the most performant GPU in the lineup - early samples have already been taken apart and reverse engineered, revealing just how unique its engineering is.

At their CES keynote earlier this month, Nvidia officially revealed the entire RTX 50XX series, though samples had likely been sent to various testing and review groups as far back as December. Among the many reviewers, the Der8auer group - who received a special "Founders edition" (FE) version of the 5090 – commented on how each of the 58 cooling fins are uniquely shaped, resulting in just two fans able to manage a device with a TDP of 575W in a FHFL PCIe form factor.

Source: Nvidia

Its counterpart in the current generation, the RTX 4090 as released in 2022, was notably difficult to acquire due to cryptocurrency miners and low-tier AI datacentres using it as a cheaper and more efficient option than datacentre grade devices which are priced for the enterprise and require warranties and licensing among other things. The RTX 5090 reference design (RD) which will be sold as the standard version is similarly expected to be extremely difficult to buy due to incredible demand.



Source: Nvidia

As for its specs (1 2 3 4):

- Using TSMCs 4nm process node
- 92.2B transistors in total within 744mm^2
- 21,760 CUDA cores
- Clocking up to 2.4GHz (2.54 for the FE)
- 32GB of GDDR7 (512-bit bus width)
- 1.79TB/s memory bandwidth
- Air-cooled TDP of 575W
- PCIe 5.0 x16 in a FHFL form factor
- 680 5th gen. Tensor cores
- Starting at ~$2000 (estimated)

Vendor spotlight:

GigaIO

A relatively small American OEM, GigaIO decided to forgo integrating network fabrics tightly with compute and memory and instead decided to extend PCIe to

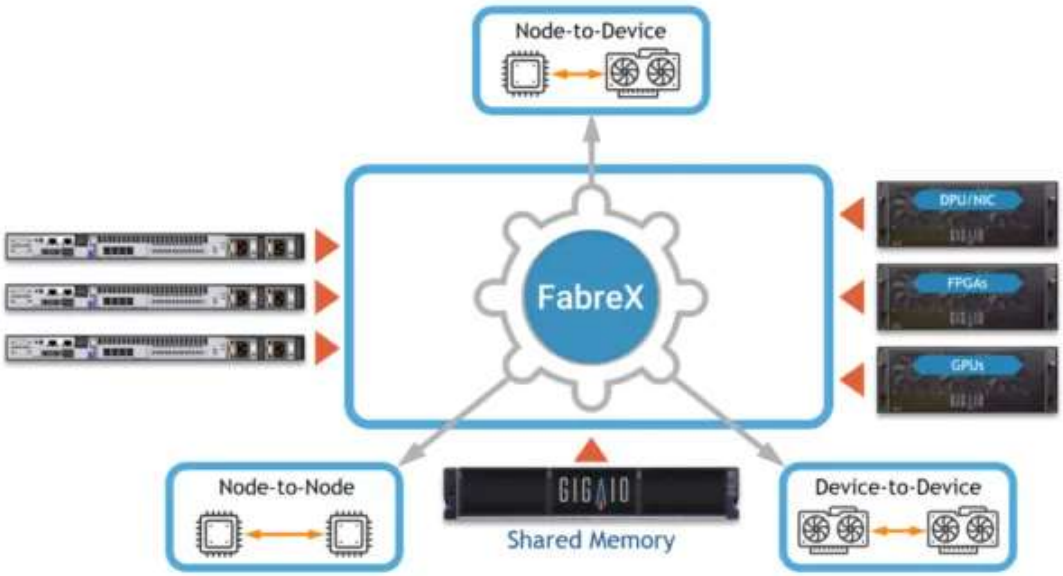
outside the chassis. With their incredibly underrated "FabreX" tech stack, they're able to connect up to 64 nodes and present them as one "giga node" to software.

Optimised for edge inferencing, the Gryf suitcase-sized supercomputer packs up to 2.5KW in a airplane form factor. Gryf comes with six field-replaceable sleds that can be customised between four categories - compute, AI accelerator, networking, and storage units. The sleds are all internally connected by a FabreX backend network as well as allowing for up to 100GbE scale out, and multiple Gryf units can be connected using FabreX as a scale-up fabric. Currently, the compute sleds support 64 core AMD Milan CPUs, and the AI accelerators that fit the size and power envelope are limited to Nvidia's L40S, though it appears other FHFL options are being qualified currently. Storage sleds have a capacity of up to 246TB, leading to well over a PB of managed storage in a Gryf optimised to act as a GPFS server.

Source: GigaIO

The FabreX networking stack essentially takes PCIe out from within the confines of a chassis and allows for composing a large amount of compute resources over many physical servers into one giant virtual machine. Technology like this does already exist but suffers from a variety of limitations and performance issues

due to the challenges involved in virtualising away many different communication protocols and networking stacks. With FabreX, everything is within the PCIe domain, and so using GigaIO's FabreX NICs and switches, everything can appear as if it is on one motherboard. This allows for aggregating large amounts of compute resources such as up to 64 GPUs into one virtual server, or for dynamically changing the setup of a machine to vary the amount of memory or networking available. This has led to a variety of bold claims that GigaIO makes on the performance of their solutions.



Source: GigaIO

One-pagers:

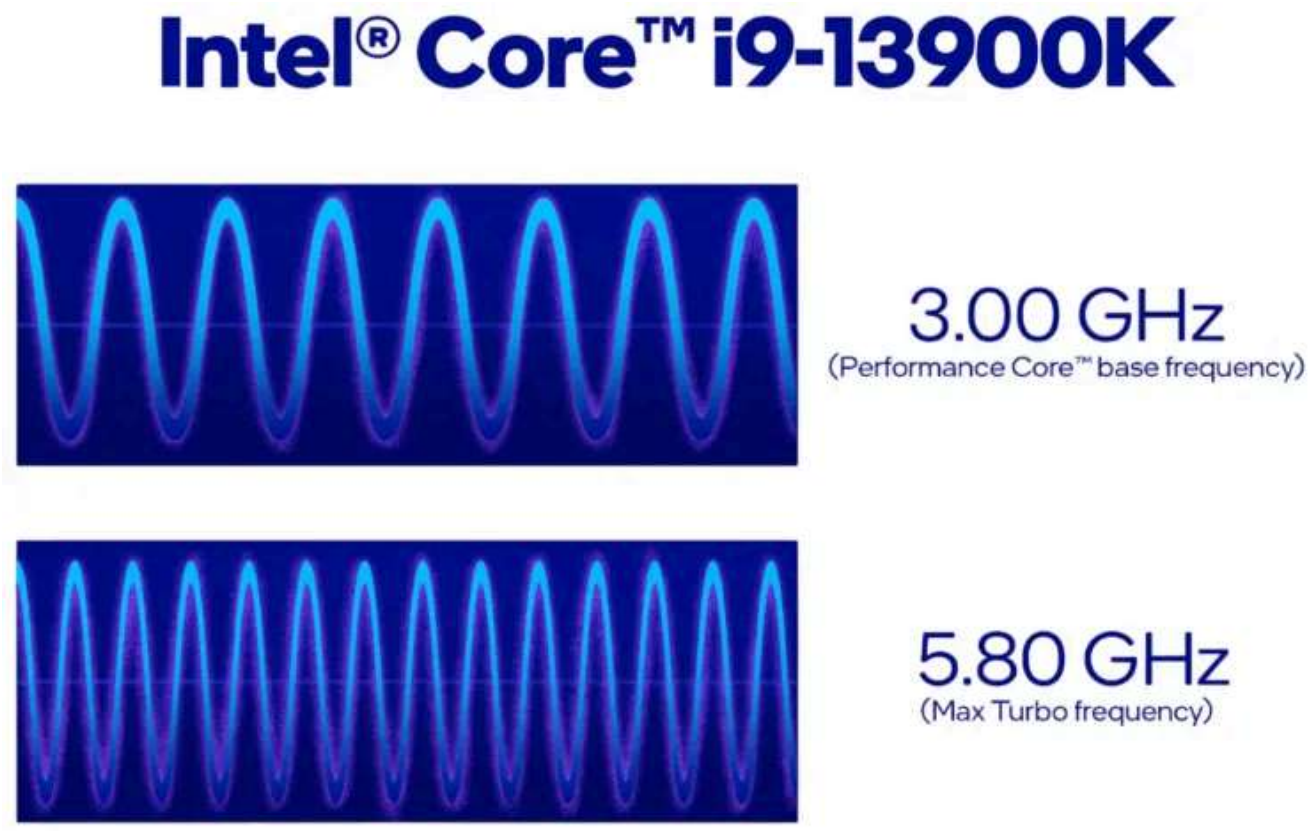
Overclocking

When power draw and cooling are no longer concerns, overclocking can offer a way to get more performance from the same hardware quickly, though the types of applications that can benefit from pushing CPUs and GPUs beyond their recommended bounds are limited.

All digital processors execute tasks by breaking them into "cycles", synchronized to an internal clock - a high-frequency signal measured in mega/gigahertz (M/GHz). Each cycle allows the processor to perform operations like fetching data, executing instructions, or writing results. Modern CPUs use techniques

such as pipelining to overlap these tasks across multiple cycles, but all operations are still tied to the clock's rhythm. For example, a CPU might complete 1–4 instructions per cycle depending on its architecture. The clock rate - essentially the speed of this internal metronome - directly influences how quickly these operations occur. Higher clock speeds reduce latency for tasks like gaming or single-threaded applications, where rapid execution of individual instructions matters most.

However, increasing the clock rate (overclocking) comes with trade-offs. While it boosts performance, it also raises power consumption and heat output, with each 10% increase in clock speed typically raises power draw by 20-30%, straining cooling systems and risking thermal throttling.



Source: Intel

Modern CPUs use dynamic frequency scaling to balance efficiency and performance, but overclocking bypasses these safeguards, prioritizing raw speed. This involves manually adjusting the CPU's clock rate and voltage via BIOS settings or software tools, often used by gamers and enthusiasts to maximize frame rates or benchmark scores. Some AI workloads, like small-scale inference,

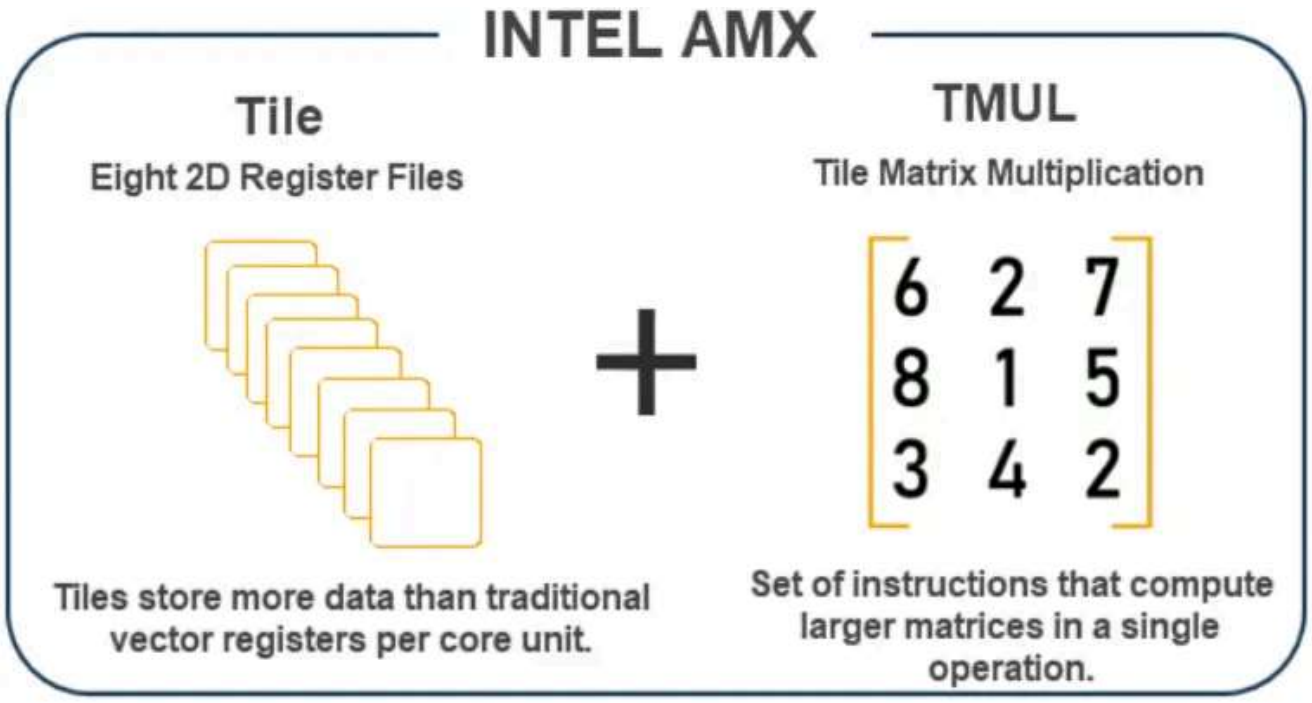
may also benefit from faster serial computations, though parallel tasks (e.g., large-scale model training) depend more on core count and memory bandwidth.

CPUs typically offer a wide variation in allowable clock rates due to the inherent flexibility of their design and having to orchestrate the various components of a server such as memory and I/O, whereas GPUs and more application-specific processors tend to have a narrower acceptable range.

Intel's AVX/AMX – Advanced Vector/Matrix eXtensions

Application performance on any processors depends not only on the inherent capability of the hardware, but also the software that runs on top of it. Intel's AVX/AMX (advanced vector/matrix extensions) are instruction sets add-ons that optimise the use of Intel CPU/NPUs for modern data-heavy workloads.

Modern CPUs rely on instruction sets - collections of commands that define how hardware processes data - to execute tasks. AVX, introduced in 2011, expanded Intel's x86 capabilities by adding 256-bit vector registers (YMM), enabling CPUs to perform operations on multiple data points, from 4 x FP64 to 32 x Int8 simultaneously. AVX accelerated tasks like image processing, physics simulations, and machine learning inference, where identical operations on large datasets benefit from parallel execution. Later, AVX-512 (2016) doubled register width to 512 bits (ZMM), further boosting throughput for scientific computing and AI training, though at higher power costs. However, with AI workloads demanding even higher arithmetic intensity, a newer instruction set with additional data structures was needed.



AMX, launched in 2023 with Intel's Sapphire Rapids CPUs, targets matrix operations by introducing 2D tiles, which are dedicated on-chip memory blocks for matrix operations. Unlike AVX's general-purpose registers, AMX tiles are optimized for matrix multiplications, a core operation in neural networks. Using tiles allow developers to further optimise processing very large regions of memory by exploiting locality and parallelism further. This design mirrors GPU-style efficiency for AI tasks while retaining CPU flexibility for mixed workloads. For example, a CPU with AMX can handle both database queries and real-time inference to a degree, making it ideal for edge AI or hybrid cloud environments.

AVX and AMX reflect Intel's strategy to bridge the gap between CPUs and specialized accelerators. AVX's vector parallelism suits diverse workloads, from video encoding to physics simulations, while AMX's matrix focus competes directly with GPUs in AI efficiency. Though AVX-512's power demands limit its adoption, AMX's lower energy footprint positions it as a scalable AI solution suitable for high-bandwidth CPUs such as Sapphire rapids onwards.

Previous Next

Write a comment...